REMARKS

Status of the Claims

Pending claims

Claims 1 and 2 as filed are pending.

Claims amended, canceled and added in the instant amendment

Claims 1 and 2 are canceled, without prejudice, and new claims 3 to 27 are added. Thus, after entry of the instant amendment, claims 3 to 27 will be pending.

Outstanding Rejections

Claims 1 and 2 stand rejected under 35 U.S.C. §112, second paragraph. Claims 1 and 2 stand rejected under 35 U.S.C. 103(a) as allegedly unpatentable over Gaasterland et al. (1998) Microbial. & Comparative Genomics 3:177-191, in view of Dandekar et al. (1998) Trends Biochem. Sci. 23:324-328 (hereinafter "Gaasterland" and "Dandekar," respectively). Applicants respectfully traverse all outstanding objections to the specification and rejections of the claims.

Support for the Claim Amendments

The specification sets forth an extensive description of the invention in the new and amended claims. Support for new claims directed to combining the Rosetta Stone method and the Phylogenetic Profile method for identifying functional links between proteins can be found, *inter alia*, on page 24, lines 8-20; page 37, lines 8-21, and claims 1and 2 as originally filed.

Support for claims directed to methods wherein a functional linkage between two proteins whose sequences are not homologous to each other can be identified by using a third protein can be found, *inter alia*, on page 3, line 26 to page 4, line 9. Support for claims drawn to methods for generating an expression profile for each protein of the genome indicating the level

of mRNA expression can be found, *inter alia*, on page 24, lines 8-10; page 37, lines 11-14; and page 38, lines 3-15.

Support for claims directed to methods for identifying proteins in a genome as being functionally linked by generating a profile for each protein in the first genome wherein the elements of the profile indicate whether a homolog of the corresponding protein is present or absent in one or more additional genomes can be found, *inter alia*, on page 19, line 8, to page 24, line 6.

Support for claims directed to methods wherein a significant sequence similarity in the alignment of sequences of (a) to the sequence of (b) is based on the degree of amino acid sequence similarity between an approximately 50 amino acid long non-homologous amino acid sequence segment of (a) to the sequence of step (b) can be found, *inter alia*, on page 14, lines 6 to 7. Support for claims directed to methods wherein a significant sequence identity in the alignment of sequences of (a) to the sequence of step (b) is based on a statistically significant alignment score can be found, *inter alia*, on page 15, lines 24 to 27. Support for claims directed to methods of identifying a "Rosetta stone" fusion protein capable of identifying functionally-linked proteins can be found, *inter alia*, in the section from page 11, line 10 to page 19, line 6 (see, e.g., page 11, lines 22 to 29), and, in the Examples page 27, line 6 to page 32. Support for claims directed to methods relying on filtering out statistically insignificant Rosetta stone links based on an excess number of homologues (paralogs) found for either protein in (a), including methods wherein an excessive number of other distinct amino acid sequence segments is present when a domain is linked to more than about 100 other domains can be found, *inter alia*, on page 8, lines 15 to 16; Figure 7; page 17, line 22 to page 18, line 10; page 32 lines 1 to 11. Support for claims directed to methods relying on filtering out statistically insignificant Rosetta Stone links when either protein in (a) forms an excessive number of Rosetta Stone links to other distinct proteins, wherein an excessive number of Rosetta Stone links is more than about 100 other domains or is more than about 25 other domains, can be found, *inter alia*, on page 17, line 27 to page 18, line 10; page 32 lines 1 to 11.

Applicants respectfully submit that no new matter has been added by the instant amendment.

Informalities

Specification drawings

The Patent Office objected to the specification because the Brief Description of the Drawings refers to a Figure 8 while the application as filed contains Fig. 8a, Fig. 8b, and Fig. 8c. Applicants' instant amendment corrects this informality and also a typographical error.

Hypertext

The Patent Office objected to the specification because the disclosure contains embedded hypertext. Applicants' instant amendment corrects this informality.

Issues under 35 U.S.C. §112, second paragraph

Claims 1 and 2 stand rejected under 35 U.S.C. §112, second paragraph. The Patent Office alleges that the phrase "sequence of multiple distinct non-homologous polypeptides" is vague and confusing. The instant amendment addresses this issue.

Issues under 35 U.S.C. §103(a)

Gaasterland in view of Dandekar

Claims 1 and 2 stand rejected under 35 U.S.C. §103(a) for allegedly being unpatentable over Gaasterland in view of Dandekar.

For a proper rejection under 35 U.S.C. §103(a), the references, either alone or in proper combination, must teach or suggest all the claim limitations of Applicants' claimed invention. Applicants will show that the deficiencies of Gaasterland are not cured by Dandekar. Accordingly, a *prima facie* case of obviousness has not been established and the rejection cannot be applied to the new claims.

Gaasterland discloses a system for carrying out cross-genome comparisons of open reading frames (ORFs) from multiple genomes. The implementation of the system includes a genome profiling system which purports to allow pairwise comparisons at different levels of match similarity and ask biologically motivated queries involving number and identity of ORFs,

their function, functional category, distribution in genomes or in biological domains, and statistics on their matches and match families.

Gaasterland uses genomic signatures compiled from comparisons of the amino acid translation of an ORF in a query genome with ORFs in a target genome. Gaasterland further describes using the genomic signature to derive summary signatures and biological domain signatures, characteristic signatures, match families and to partition ORFs into functional categories. As stated throughout the reference, Gaasterland looks to compare multiple microbial genomes by analyzing complex relationships among ORFs. One such relationship was the assignment of a biological domain class to every ORF that summarized its matches in each domain. Another such relationship was the identification of ORFs that are common to all members of a set of genomes, but absent from all others. Again, the analysis is directed to comparing multiple genomes. The signatures form a dataset for deriving statistical inferences about sets of genomes, for example, assessing monophyly by evaluating whether a match in one genome increases the frequency of a match in another genome.

In contrast, Applicants' claimed invention is directed towards methods for identifying functional links between proteins. Independent claims 3, 9, and 19 utilize both a Rosetta Stone method and a Phylogenetic Profile method to identify functionally linked proteins.

The present invention requires the use of the Rosetta Stone method for identifying functionally related proteins. Neither Gaasterland nor Dandeker teaches or suggests the Rosetta Stone method.

The Rosetta Stone method includes the step of aligning a sequence from non-homologous first and second proteins to a third protein. The Phylogenetic Profile method includes the step of grouping together proteins having similar phylogenetic profiles, wherein the similar profiles indicate that the proteins are functionally linked.

While Gaasterland analyzes multiple genomes, Applicants' disclosure analyzes proteins and functional links between proteins. Accordingly, Gaasterland does not teach or suggest the step of aligning the sequences of two non-homologous proteins to a third protein, nor does it teach or suggest the step of indicating proteins that are functionally linked based on similar phylogenetic profiles. Thus, Gaasterland does not teach or suggest the Rosetta Stone

method, nor does it teach or suggest the Phylogenetic Profile method; therefore, it cannot teach or suggest the combination of the two methods.

The Patent Office alleges that "Gaasterland et al. place the ORFs encoding the amino acid sequences analyzed in diagrams/Tables such that functionally linked proteins are closer together and identify ORFs that fall in a group (see Tables 4 and 6)."[1] Applicants respectfully disagree with this interpretation of Gaasterland. Gaasterland does not show functionally linked proteins grouped closer together. Table 4 shows the distribution, by functional categories (described on page 182 of Gaasterland), of all ORFs in each genome that find a level 3 or better match with an ORF in another genome.[2] Table 4 does not show functionally linked proteins grouped closer together. Applicants submit that while an ORF having a level 3 or better match with an ORF in another genome is placed into one of 15 functional categories in Table 4, it does not show which of the proteins of these ORFs are functionally linked. Similarly, Table 6 looks to see whether, in each genome, functional categories are overrepresented or underrepresented in each domain signature class at each level.[3] Table 6 also does not show which of the proteins of these ORFs are functionally linked.

One difference between the "genomic signatures" described by Gaasterland and the "phylogenetic profiles" of the claimed invention is in their use. The instant invention describes a method for using phylogenetic profiles to identify functionally linked proteins: proteins that participate in the same cellular process and are potentially interacting. These functionally linked proteins do not need to be homologous. Therefore, instant claimed methods represent a key departure from previous methods, which group genes and proteins by sequence similarity. The instant invention found that proteins that are not homologous, but have similar phylogenetic profiles, are likely to participate in the same cellular process(es) and are therefore likely to be functionally linked. Therefore, the claimed method can be used to identify the general function of proteins that are not homologous to any functionally characterized proteins. Gaasterland does not infer functions of proteins that are not homologous.

---

1 See page 4, lines 13-15, of the Office Action.
2 See page 184, fifth line to fourth line from the bottom, of Gaasterland.
3 See page 185, second line form the bottom, to page 186, first line under Table 3, of Gaasterland.

In further contrast to the claimed invention, the genomic signatures described by Gaasterland are used to determine how cellular functions are distributed across phyla. Gaasterland shows in tables 4 and 6 how functional categories are distributed across the organisms with fully sequenced genomes.

This method is distinctly different from that of the claimed invention. Nowhere within Gaasterland is it proposed that genes or proteins with similar genomic signatures are likely to act within the same cellular process. Furthermore, nowhere within Gaasterland does it propose that genes or proteins that cannot be characterized by virtue of their sequence similarity to characterized genes may be assigned a function using the genomic signatures.

Dandekar is cited to cure the deficiencies of Gaasterland. Dandekar discloses a systematic comparison of genomes that revealed a low level of gene-order (and operon architecture) conservation. Ordered gene pairs that were conserved appeared to interact physically. Danderkar's catalog of the local order of genes is distinctly different from the catalog of the presence and absence of genes of the Phylogenetic Profile, and Gaasterland's genomic signatures.

Dandekar does <u>not</u> cure the deficiencies of Gaasterland and, therefore, its teachings cannot be combined with the teachings of Gaasterland to arrive at Applicants' claimed invention. For example, like Gaasterland, Dandekar does not teach or suggest the step of aligning the sequences of two non-homologous proteins to a third protein of the Rosetta Stone method, nor does it teach or suggest the step of indicating which proteins are functionally linked based on similar phylogenetic profiles of the Phylogenetic Profile method. In fact, Dandekar teaches a completely different and unrelated method for identifying proteins that interact together, *i.e.*, gene order conservation, which does not even suggest looking at the presence or absence of homologous proteins across genomes to identify functionally linked proteins.

As neither Gaasterland nor Dandekar, alone or in proper combination, teaches or suggests the Rosetta Stone method or the Phylogenetic Profile method for identifying functional links between proteins, they cannot teach the combination of the Rosetta Stone method with the Phylogenetic Profile method for identifying functional links between proteins. They do not teach or suggest all the claim limitations of Applicants' claimed invention.

Applicants further aver that there is no suggestion or motivation in Gaasterland to combine it with the teachings of Dandekar to arrive at the claimed invention. Dandekar et al. describes the "Gene Neighbor" method of detecting interacting proteins from fully sequenced genomes. Dandekar does not describe the "Phylogenetic Profile" method of the claimed invention, which is based on a different assumption and which yields different, often complementary, information to the Gene Neighbor method.

The central assumption of Dandekar's Gene Neighbor method is that if the order of adjacent genes in two or more genomes is conserved, then the encoded proteins are physically interacting. Dandekar clearly states in numerous places that they are focusing on order of adjacent, or near adjacent genes. The central importance of ORDER is given in Dandekar's title, in the abstract, in the first paragraph, in the first sentence of the second paragraph, in the first sentence of the third paragraph, in the subtitle at the start of the fourth paragraph, and elsewhere.

The central assumption of the Phylogenetic Profile method of the invention is that if the pattern of presence or absence in two or more genomes is conserved for two proteins, then these proteins are functionally linked (i.e., participate in the same metabolic or signaling pathway, or interact physically). There is no assumption about the order of the genes and there is no assumption about the proximity of genes, as there is in the gene neighbor method of Dandekar. Also, the correlated absence of two encoded proteins from a genome in the Phylogenetic Profile method of the claimed invention is just as important as the correlated presence.

The key distinction between the claimed methods and Dandekar is that Dandekar's "Gene neighbor" method measures relative positions of genes on a genome while the Phylogenetic Profiles method of the claimed invention measures the presence or absence of genes in a genome, irrespective of their position. Although both methods may be used to generate functional links between genes, the methods use different approaches to generate these links.

In view of the above remarks, Applicants submit that the pending claimed invention is distinguished and not rendered obvious by the cited art. Accordingly, the Examiner is respectfully requested to withdraw the rejection under 35 U.S.C. §103(a).

CONCLUSION

In view of the foregoing amendment and remarks, it is believed that the Examiner can properly withdraw the rejection of the pending claims under 35 U.S.C. §112, second paragraph and 35 U.S.C. §103(a). Applicants believe all claims pending in this application are in condition for allowance. The issuance of a formal Notice of Allowance at an early date is respectfully requested.

Attached is a marked-up version of the changes being made by the current amendment.

If necessary, please apply any additional and necessary charges, and apply all credits, to Deposit Account No. 06-1050.

If the Examiner believes a telephone conference would expedite prosecution of this application, please telephone the undersigned at (858) 678-5070.

Respectfully submitted,

Date: July 15, 2002

44,830

Gregory P. Einhorn
Reg. No. 38,440

Fish & Richardson P.C.
4350 La Jolla Village Drive, Suite 500
San Diego, California 92122
Telephone: (858) 678-5070
Facsimile: (858) 678-5099

VERSION WITH MARKINGS TO SHOW CHANGES MADE

Applicant : Marcotte et al.
Serial No. : 09/493,401
Filed : January 28, 2000
Title : COMBINED COMPUTATIONAL METHODS FOR DETECTING PROTEIN FUNCTION AND PROTEIN-PROTEIN INTERACTIONS FROM GENOME SEQUENCES

Art Unit : 1631
Examiner : Shubo Zhou, Ph.D.

The above-captioned application has been amended as follows:

In the Specification:

The paragraph beginning on page 8, line17, has been amended as follows:

FIGs. 8A-8C are diagrams [FIG 8 is a diagram] showing the process and result of the method of phylogenetic profiles. In each case all proteins with identical profiles to the query proteins were found (within the double box) and then all those with profiles that differed by one bit were found (in the second column). Proteins in bold face participate in the same complex or pathway as the query protein and in italics participate in a different but related complex or pathway. Proteins with identical profiles are shown within a box. Single lines between boxes represent a one-bit difference between the two profiles. All neighboring proteins whose profiles differ by one bit from the query protein are shown. Homologous proteins are connected by a dashed line or indented. Each protein is labeled by a four-digit *E. coli* number, a Swissprot gene name and a brief description. Notice that proteins within a box or in boxes connected by a line have similar functions. Hypothetical proteins (*i.e.* of unknown function) are prime candidates for functional and structural studies. Proteins in the double boxes in 8(a), 8(b) and 8(c) have, respectively, 11, 6, and 10 ones in their phylogenetic profiles, out of a possible 16 for the 17 genomes available at the time of calculation.

The paragraph beginning on page 2, line 12, has been amended as follows:

In order to more fully understand and determine potential therapeutics, antibiotic and biologics for various organisms, efforts have been taken to sequence the genomes of a number of organisms. For example the Human Genome Project began with the specific goal of obtaining the complete sequence of the human genome and determining the biochemical function(s) of each gene. To date, the project has resulted in sequencing a substantial portion of the human genome (J. Roach, on the website of the University of Washington [http://weber.u.Washington.edu/~roach/human_ genome_progress2.html]) (Gibbs, 1995). At least twenty-one other genomes have already been sequenced, including, for example, *M. genitalium* (Fraser *et al.*, 1995), *M. jannaschii* (Bult *et al.*, 1996), *H. influenzae* (Fleischmann *et al.*, 1995), *E. coli* (Blattner *et al.*, 1997), and yeast (*S. cerevisiae*) (Mewes *et al.*, 1997). Significant progress has also been made in sequencing the genomes of model organism, such as mouse, *C. elegans*, *Arabadopsis sp.* and *D. melanogaster*. Several databases containing genomic information annotated with some functional information are maintained by different organization, and are accessible via the internet, for example, the websites of the Institute for Genomic Research [http://wwwtigr.org/tdb]; the University of Wisconsin Laboratory for Genetics [http://www.genetics.wisc.edu]; Stanford University's Dept. of Genetics [http://genome-www.stanford.edu/~ball]; the Los Alamos National Laboratories HIV databases [http://hiv-web.lanl.gov]; the National Center for Biotechnology Information [http://www.ncbi.nlm.nih.gov]; the European Bioinformatics Institute [http://www.ebi.ac.uk]; the Institut Pasteur Bio Netbook [http://Pasteur.fr/other/biology]; and the Whitehead Institute/MIT Center for Genome Research [http://www.genome.wi.mit.edu]. The raw nucleic acid sequences in a genome can be converted by one of a number of available algorithms to the amino acid sequences of proteins, which carry out the vast array of processes in a cell. Unfortunately, these raw protein sequence data do not immediately describe how the proteins function in the cell. Understanding the details of various cellular processes (*e.g.*, metabolic pathways,

signaling between molecules, cell division, *etc.*) and which proteins carry out which processes, is a central goal in modern cell biology.

The paragraph beginning on page 28, line 23, has been amended as follows:

The second test of the interactions predicted by the Rosetta Stone method uses as confirmation the Database of Interacting Proteins provided at the website of the UCLA DOE laboratory [(http://doe-mbi.ucla.edu)]. This is a compilation of protein pairs that have been found to interact in some published experiment. As of December 1998, the database contained 939 entries, 724 of which have both members of the pair listed in the ProDom database. Of these 724 pairs, we find 46 or 6.4% linked by Rosetta Stone sequences. We expect this percentage to rise as more genomes are sequenced, revealing more linked sequences.

In The Claims:

Claims 1 and 2 have been canceled, without prejudice.

The following new claims have been added:

--3. (NEW) A method for identifying a high confidence functional link between at least two proteins, comprising the following steps:

(a) identifying non-homologous proteins as being functionally linked by a "Rosetta Stone" method comprising the following steps

(i) providing amino acid sequences of a first protein and a second protein, wherein the first and second proteins are not homologous,

(ii) providing an amino acid sequence of a third protein,

(iii) aligning amino acid sequence segments from the first protein and the second protein to the amino acid sequence of the third protein, wherein the amino acid sequence segments from the first and the second protein do not align to each other with any significant sequence similarity, and

(iv) establishing whether the first and second proteins are functionally linked by determining whether a significant sequence similarity is present between the

aligned amino acid sequences of step (iii), thereby identifying non-homologous proteins as being functionally linked;

(b) identifying pairs of proteins in a genome as being functionally linked by a "phylogenetic profile" method comprising the following steps

(i) providing a first plurality of protein sequences comprising substantially all protein sequences encoded by a first genome,

(ii) providing a second plurality of protein sequences comprising substantially all protein sequences encoded by one or more additional genomes,

(iii) comparing each protein sequence in the first plurality of protein sequences with substantially all the protein sequences of the second plurality of protein sequences to determine if a protein sequence in the first genome has a homolog in the one or more additional genomes based on the degree of similarity of the sequences being compared,

(iv) generating a phylogenetic profile for each protein of the first genome, wherein the phylogenetic profile is a vector or pattern whose elements indicate whether a homolog of the corresponding protein is present or absent in the one or more additional genomes, and

(v) grouping together proteins having similar phylogenetic profiles, wherein a similar phylogenetic profile indicates a functional link between the proteins; and

(c) identifying pairs of proteins that are linked in both (a) and (b), thereby identifying a high confidence functional link between at least two proteins.

4. (NEW) The method of claim 3, further comprising:

generating an expression profile for each protein of the genome where the expression profile is a vector or a pattern whose elements indicate the level of mRNA expression of the corresponding gene in two or more DNA chip experiments; and

grouping together genes having similar expression profiles where a similar expression profile indicates a functional link between proteins.

5. (NEW) The method of claim 4, further comprising displaying the functional links as networks of related proteins, comprising:

placing a plurality of proteins in a diagram such that functionally linked proteins are closer together than all other proteins; and

identifying groups of proteins that fall in a cluster in said diagram as functionally related.

6. (NEW) The method of claim 5, wherein the placing of the plurality of proteins in a diagram utilizes a computer.

7. (NEW) The method of claim 3, further comprising:

identifying functional links for a plurality of protein pairs;

placing substantially all protein pairs that are identified as functionally linked in a diagram such that functionally linked proteins are closer together than other proteins; and

identifying groups of proteins that fall in a cluster in said diagram as functionally related.

8. (NEW) The method of claim 7, wherein the placing of substantially all protein pairs in a diagram utilizes a computer.

9. (NEW) A method for identifying a high confidence functional link between at least two proteins, comprising the following steps:

(a) identifying non-homologous proteins as being functionally linked by a "Rosetta Stone" method comprising the following steps

(i) providing a pair of non-homologous protein amino acid sequences;

(ii) providing an amino acid sequence of a third protein;

(iii) aligning amino acid sequence segments from the first protein in (i) and the second protein in (i) to the amino acid sequence of (ii) where the first and the second protein of (i) are not homologues; and

(iv) establishing whether a significant sequence similarity is present between the alignments of (iii), wherein identification of a significant sequence similarity between each of the two non-homologous amino acid sequence segments from two proteins of (i) to different sequence segments of the protein of (ii) identifies the pair of proteins of (i) as being functionally linked each other;

(b) identifying pairs of proteins in a genome as being functionally linked by a "phylogenetic profile" method comprising the following steps

(i) providing a first plurality of protein sequences comprising substantially all protein sequences encoded by a first genome,

(ii) providing a second plurality of protein sequences comprising substantially all protein sequences encoded by one or more additional genomes,

(iii) comparing each protein sequence in the first plurality of protein sequences with substantially all the protein sequences of the second plurality of protein sequences to determine if a protein sequence in the first genome has a homolog in the one or more additional genomes based on the degree of similarity of the sequences being compared,

(iv) generating a phylogenetic profile for each protein of the first genome, wherein the phylogenetic profile is a vector or pattern whose elements indicate whether a homolog of the corresponding protein is present or absent in the one or more additional genomes, and

(v) grouping together proteins having similar phylogenetic profiles, wherein a similar phylogenetic profile indicates a functional link between the proteins; and

(c) identifying pairs of proteins that are linked in both (a) and (b), thereby identifying a functional link between at least two proteins.

10. (NEW) The method of claim 9, wherein in the "Rosetta Stone" method establishing that the pair of non-homologous amino acid sequence segments of (i) have significant sequence similarities to different sequence segments of the protein of (ii) comprises showing that a computed probability (p) value is below a statistically significant threshold.

11. (NEW) The method of claim 10, wherein the probability threshold is set with respect to a value 1/N, wherein N is an integer based on the total number of protein sequences in a database.

12. (NEW) The method of claim 9, wherein in the "Rosetta Stone" method the non-homologous amino acid sequence segments from different protein sequences of (i) are at least about 50 amino acid residues long.

13. (NEW) The method of claim 9, wherein in the "Rosetta Stone" method the non-homologous amino acid sequence segments from different polypeptide sequences of (i) are between about 50 and about 1000 amino acid residues long.

14. (NEW) The method of claim 9, wherein in the "Rosetta Stone" method statistically insignificant Rosetta stone links are filtered out when either protein in (i) has a plurality of homologs.

15. (NEW) The method of claim 9, wherein in the "Rosetta Stone" method the plurality of homologues is more than about 100 homologues.

16. (NEW) The method of claim 9, wherein in the "Rosetta Stone" method statistically insignificant Rosetta Stone links are filtered out when either protein in (i) forms a plurality of Rosetta Stone links to other distinct proteins.

17. (NEW) The method of claim 16, wherein the plurality of Rosetta Stone links is more than about 100.

18. (NEW) The method of claim 17, wherein the plurality of Rosetta Stone links is more than about 25.

19. (NEW) A method for identifying a high confidence functional link between at least two proteins, comprising the following steps:

(a) identifying non-homologous proteins as being functionally linked by a "Rosetta Stone" method comprising the following steps

(i) providing amino acid sequences of a first protein and a second protein, wherein the first and second proteins are not homologous,

(ii) providing an amino acid sequence of a third protein,

(iii) aligning amino acid sequence segments from the first protein and the second protein to the amino acid sequence of the third protein, wherein the amino acid sequence segments from the first and the second protein do not align to each other with any significant sequence similarity, and

(iv) establishing whether the first and second proteins are functionally linked by determining whether a significant sequence similarity is present between the aligned amino acid sequences of step (iii), thereby identifying non-homologous proteins as being functionally linked;

(b) identifying pairs of proteins in a genome as being functionally linked by a "phylogenetic profile" method comprising the following steps

(i) providing a first plurality of protein sequences comprising substantially all protein sequences encoded by a first genome, or, a plurality of nucleic acid sequences comprising substantially all protein-encoding nucleic acid sequences in a first genome;

(ii) providing a second plurality of protein sequences comprising substantially all protein sequences encoded by one or more additional genomes, or, a

second plurality of nucleic acid sequences comprising substantially all protein-encoding nucleic acid sequences of one or more additional genomes;

(iii) comparing each protein sequence or nucleic acid sequence in the first plurality of protein sequences or nucleic acid sequences respectively with substantially all the protein sequences or nucleic acid sequences of the second plurality of protein sequences or nucleic acid sequences to determine if the protein sequence or nucleic acid sequence in the first genome has a homolog in the one or more additional genomes based on the degree of similarity of the sequences being compared;

(iv) generating a phylogenetic profile for each protein of the first genome, wherein the phylogenetic profile is a vector or pattern whose elements indicate whether a homolog of the corresponding protein or nucleic acid is present or absent in the one or more additional genomes; and

(v) grouping together proteins having similar phylogenetic profiles, wherein proteins with similar profiles are identified as being functionally linked; and

(c) identifying pairs of proteins that are linked in both (a) and (b), thereby identifying a functional link between at least two proteins.

20. (NEW) The method of claim 19, wherein the phylogenetic profile is generated using a bit type profiling method.

21. (NEW) The method of claim 19, wherein the phylogenetic profile is generated using an evolutionary distance method.

22. (NEW) The method of claim 19, wherein the phylogenetic profile is generated in a binary code describing the presence or absence of a given protein in an organism.

23. (NEW) The method of claim 19, wherein the phylogenetic profile is generated in a continuous code that describes how similar the related sequences are in the different genomes.

24. (NEW) The method of claim 19, wherein the phylogenetic profile is generated using an evolution probability process, wherein the process comprises

(a) constructing a conditional probability matrix: p(aa → aa'), where aa and aa' are any amino acids, and the conditional probability matrix is constructed by converting an amino acid substitution matrix from a log odds matrix to a conditional probability matrix;

(b) accounting for an observed alignment of the constructed conditional probability matrix by taking the product and the conditional probabilities for each aligned pair of amino acids during the alignment of the two protein sequences, represented by

$$P(p)=\prod_{n} p(aa_n \rightarrow aa'_n)\,;\ \text{and}$$

(c) determining an evolutionary distance α from powers equation:

$p'=p^{\alpha}$ (aa → aa'), maximizing for P.

25. (NEW) The method of claim 24, wherein the conditional probability matrix is defined by a Markov process with substitution rates over a fixed time interval.

26. (NEW) The method of claim 24, wherein the conversion from an amino acid substitution log odds matrix to a conditional probability matrix is represented by:

$P_{\beta}(i \rightarrow j) = P(j)2^{\wedge}$ [BLOSUM62 ij / 2] ,

where BLOSUM62 is an amino acid substitution log odds matrix, and P(i→j) is the probability that amino acid i is replaced by amino acid j through point mutations according to BLOSUM62 scores.

27. (NEW) The method of claim 26, wherein Pj's are the abundances of amino acid j and are computed by solving a plurality of linear equations given by the normalization condition that $\sum_i P_B(i \rightarrow j) = 1$.--